Subspace Adversarial Training

Tao Li Yingwen Wu Sizhe Chen Kun Fang Xiaolin Huang
Department of Automation, Shanghai Jiao Tong University
{li.tao, yingwen_wu, sizhe.chen, fanghenshao, xiaolinhuang}@sjtu.edu.cn

Abstract

*Single-step adversarial training (AT) has received wide attention as it proved to be both efficient and robust. However, a serious problem of catastrophic overfitting exists, i.e., the robust accuracy against projected gradient descent (PGD) attack suddenly drops to 0% during the training. In this paper, we approach this problem from a novel perspective of optimization and firstly reveal the close link between the fast-growing gradient of each sample and overfitting, which can also be applied to understand robust overfitting in multi-step AT. To control the growth of the gradient, we propose a new AT method, **Sub**space **A**dversarial **T**raining (**Sub-AT**), which constrains AT in a carefully extracted subspace. It successfully resolves both kinds of overfitting and significantly boosts the robustness. In subspace, we also allow single-step AT with larger steps and larger radius, further improving the robustness performance. As a result, we achieve state-of-the-art single-step AT performance. Without any regularization term, our single-step AT can reach over $\mathbf{51}\%$ robust accuracy against strong PGD-50 attack of radius $8/255$ on CIFAR-10, reaching a competitive performance against standard multi-step PGD-10 AT with huge computational advantages. The code is released at https://github.com/nblt/Sub-AT.*

1. Introduction

Adversarial training (AT) [23], which aims to minimize the model's risk under the worst-case perturbations, is currently the most effective approach for improving the robustness of deep neural networks. For a given neural network $f(\mathbf{x}, \mathbf{w})$ with parameters $\mathbf{w}$, the optimization objective of AT can be formulated as follows:

$$\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{B}(\mathbf{x},\epsilon)} \mathcal{L}\left(f(\mathbf{x}+\delta, \mathbf{w}), y\right) \right],$$

where $\mathcal{B}(\mathbf{x}, \epsilon)$ is the norm ball with radius ϵ and $\mathcal{L}$ is the loss function. The key issue of AT lies in solving the inner worst-case problem by generating adversarial examples.

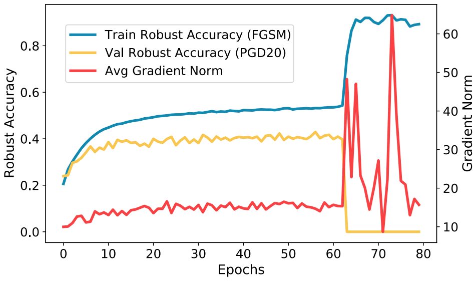

Figure 1. *Catastrophic overfitting* in single-step AT. The experiments are conducted on CIFAR-10 with PreAct ResNet18 model for adversarial robustness against ℓ_∞ perturbations of radius $8/255$. The robust accuracy of single-step Fast AT on the validation set against PGD-20 attack abruptly drops to 0 in one single epoch, characterized by a rapid explosion of the average gradient norm of each sample.

Presently the most efficient way to generate adversarial examples is the fast gradient sign method (FGSM) [9], i.e.,

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sgn}\left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}, \mathbf{w}), y)\right).$$

Since the adversarial examples above are generated by one-step gradient propagation, the corresponding AT is called *single-step* AT. In Fig. 1, we demonstrate a standard single-step AT process where the training robust accuracy against FGSM attack keeps increasing. However, the generalization capability, i.e., the robust accuracy on the validation set under projected gradient descent (PGD) attack [23], can suddenly drop to zero, which is a typical overfitting phenomenon referred as *catastrophic overfitting* [40].

Many works [1, 16, 17, 32, 37, 40] are devoted to resolving such an intriguing overfitting problem. One approach to tackle the overfitting is to use a judiciously designed learning rate schedule as well as appropriate regularizations. For example, Wong *et al.* [40] proposed to add a random step to FGSM and introduce cyclic learning rates [30] to overcome the overfitting. Andriushchenko *et al.* [1] proposed a novel regularization term called GradAlign to further improve the quality of single-step AT solutions. However, these methods highly rely on specifically designed learning rate schedules, which need to be tuned carefully for different tasks. Another approach is to generate more precise adversarial examples. For example, Kim *et al.* [17] suggested verifying the inner interval along the adversarial direction and searching for appropriate step size. PGD AT, a typical *multi-step* AT which generates adversarial examples using multiple iterations, can also help avoid catastrophic overfitting. However, these methods require multiple forward propagations. More seriously, overfitting can still prominently occur in multi-step AT (known as *robust overfitting*) as demonstrated by Rice *et al.* [28].

In order to understand this interesting phenomenon, let us investigate what happens at the 64-th epoch in Fig. 1 when catastrophic overfitting occurs. Before the overfitting, the training robust accuracy has already stepped into a stable stage, indicating the small norm of batch gradient $\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathbf{w}}\mathcal{L}(f(\mathbf{x}_i^{\mathrm{adv}},\mathbf{w}),y)\right\|_2$ (n denotes the batch size). There are two possibilities for the small batch gradient: ***i***) the gradient of each sample is small; ***ii***) the gradients of samples does not converge, but they cancel each other, resulting in an overall balanced state. We then plot the average norm of each sample's gradient on one fixed training batch (i.e., $\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\mathbf{w}}\mathcal{L}(f(\mathbf{x}_i^{\mathrm{adv}},\mathbf{w}),y)\right\|_2$) in red. An interesting thing is that before the overfitting, the average norm stays almost constant. However, it abruptly increases in the moment when the overfitting occurs. Intuitively, at that time, the balance of gradient is broken — the network tries to capture each sample's label with huge fluctuations, namely large gradients, a significant signal of overfitting. This phenomenon also coincides with the recent discussion on the connection between the gradient variance and generalization capability [12, 15, 25].

Inspired by the link between large gradients and overfitting, we propose to resolve the overfitting by controlling the magnitude of the gradient. A possible way is to restrict the gradient descent in a subspace instead of the whole parameter space, to prevent the excessive growth of the gradient. The key challenge lies in keeping the network's capability in such a subspace, which has been recently discussed in [20] showing that, optimizing parameters in a tiny subspace extracted from training dynamics could keep the performance. Based on this discovery, we propose a new AT method called ***Sub****space* ***A****dversarial* ***T****raining* (***Sub-AT***), which identifies such an effective subspace and conducts AT in it. From the training statistics of Sub-AT in Fig. 2a, we observe that it successfully controls the average gradient norm under a low level (the yellow dotted curve), thus resolving the catastrophic overfitting. Meanwhile, the robust accuracy is significantly improved from 0.4 to nearly 0.5 (the yellow solid curve). The sensitivity to learning rates is also fundamentally overcome as we only use a constant learning rate, and the results remain similar across a wide range of choices. As a direct extension, Sub-AT can be applied to mitigate the robust overfitting (Fig. 2b) in multi-step AT, implying the similar essence behind these two phenomena. Thus for the first time, the two overfittings, which were previously treated separately [1], are now connected and resolved in a unified approach.

Since training in subspace controls the gradient magnitude and hence fundamentally resolves the catastrophic overfitting, we now can allow larger steps and radius, which previously requires the assistance of delicate regularizations, e.g. GradAlign [1]. It brings further improvement on robustness, from which it follows that pure single-step-based AT (without regularization terms) achieves competitive robustness with standard multi-step PGD AT with great computational benefits, answering a long-existing question:

Can single-step AT achieve comparable robustness against iterative attacks than multi-step AT?

Our Sub-AT uncovers the long-neglected potential of single-step AT and can enlighten more efficient and powerful AT algorithms.

Our main contributions can be summarized as follows:

- We approach the *catastrophic overfitting* in single-step AT from a novel view of optimization and firstly reveal the close link between the fast-growing gradient of each sample and overfitting, which can also be applied to explain the *robust overfitting* in multi-step AT.
- We propose an efficient AT method, Sub-AT, which constrains AT in a carefully extracted subspace, to control the growth of gradient. It uniformly resolves both kinds of overfitting, significantly improves the robustness, and successfully overcomes the sensitivity to learning rates. It is also very easy to combine with other AT methods to bring consistent improvements.
- Our Sub-AT achieves *state-of-the-art* adversarial robustness on single-step AT and can successfully train with larger steps and larger radius, which brings further improvements. Notably, our pure single-step AT achieves over 51% robust accuracy against PGD-50 attack of $\epsilon = 8/255$ on CIFAR-10, competitive to the multi-step PGD-10 AT with great time benefits.

2. Related Work

Adversarial Training. Since deep neural networks are easily fooled by adversarial examples, many defense methods [4, 5, 8, 21–24, 27, 31, 33, 38, 39, 41, 43–45] have been proposed. Among them, AT [23], which augments the training data with adversarial perturbations, is currently the most effective way to improve the robustness of the model. According to the number of times the gradient propagation involved in generating adversarial perturbations, AT